ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Laurea in Ingegneria Informatica
ALLEGATO - Abstract e Indice in Italiano

Exploring Phase-Change Memory as a Compute Accelerator Module:
A Virtual Prototype on the PULP Platform

Relatore:
Chiar.mo Prof.
Giuseppe Tagliavini

Presentata da:
Leonardo Domenicali

Invernale
2024/2025

Exploring Phase-Change Memory as a Compute Accelerator Module:

A Virtual Prototype on the PULP Platform

Leonardo Domenicali

Abstract

Le architetture di Von Neumann stanno raggiungendo il limite. La separazione fisica tra memoria e unità di elaborazione crea un collo di bottiglia che sta diventando critico per le applicazioni basate sull'elaborazione di grandi quantità di dati, in particolare nell'ambito dell'Intelligenza Artificiale e del Machine Learning. Lo spostamento continuo dei dati tra CPU e memoria sta diventando più dispendioso in termini di tempo ed energia rispetto alla computazione stessa: questo è il problema noto come "*memory wall*".

Questa tesi esplora la memoria a cambiamento di fase (PCM, Phase-Change Memory) come potenziale soluzione attraverso il paradigma del calcolo analogico in memoria (AIMC, Analog In-Memory Computing). L'idea fondamentale è quella di eseguire le operazioni di calcolo direttamente dove i dati sono memorizzati, invece di affidarsi a costosi cicli di trasferimento. La PCM risulta particolarmente interessante per questo tipo di applicazioni poiché i suoi stati di resistenza analogici possono rappresentare in modo naturale i pesi di una rete neurale, e la sua architettura a matrice (crossbar) permette l'esecuzione parallela di moltiplicazioni matrice-vettore in un'unica operazione analogica.

Il principale contributo di questo lavoro è la realizzazione di un prototipo virtuale funzionante di un acceleratore AIMC basato su PCM. Il modello è implementato in C++ e integrato nella piattaforma di simulazione GVSoC, all'interno dell'ecosistema PULP. Una parte significativa del lavoro si è concentrata sull'ottimizzazione della simulazione, rendendola sufficientemente veloce da essere utile per esperimenti pratici. Ciò ha comportato la progettazione di strutture dati ottimizzate per la cache e l'implementazione di esecuzione multithread per l'algoritmo principale di moltiplicazione matrice-vettore. I test prestazionali dimostrano l'efficacia di queste ottimizzazioni: la struttura dati a buffer lineare riduce i cache miss fino a un fattore di X rispetto all'approccio tradizionale basato su array multidimensionali, mentre il multithreading offre un notevole incremento di velocità sui sistemi multicore, con prestazioni ottimali ottenute con 8 thread nelle configurazioni testate.

Il risultato è un framework di simulazione configurabile, utilizzabile per esplorare architetture AIMC e fornire una base per la co-progettazione di acceleratori hardware e del software necessario al loro utilizzo efficace.

Chapter 1

Introduzione

(pagina 1)

1.1 Il Muro della Memoria e l'architettura Von Neumann

(pagina 1)

1.2 Obbiettivi e motivazioni della tesi

(pagina 1)

1.3 Sommario dei contributi

(pagina 2)

Chapter 2

Contesto e Tecnologie Abilitanti

Chapter 3

Implementazione Modulo PCM

3.7 Integrazione in GVSoC

Chapter 4

Test e Risultati

(pagina 17)

4.1 Analisi e Test dell'Algoritmo

(pagina 17)

4.1.1 Valutazione Dati

(pagina 17)

Metriche di Prestazione

(pagina 17)

4.1.2 Analisi delle Performance

(pagina 17)

Impatto delle Ottimizzazioni del Compilatore

(pagina 19)

Comportamento delle Cache

(pagina 19)

Pattern di Accesso alla Memoria

(pagina 20)

Performance del Multithreading

(pagina 20)

Chapter 5

Conclusione e Lavori Futuri

(pagina 23)

5.1 Conclusioni

(pagina 23)

5.2 Lavori Futuri

(pagina 23)